

DOCUMENT RESUME

ED 318 747

TM 014 701

AUTHOR DeMauro, Gerald E.
TITLE Effects of Representation of Gender Groups in the Examinee Population on the Mantel-Haenszel Procedure.
PUB DATE 12 Apr 90
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Students; *Difficulty Level; Higher Education; *Item Bias; Ratios (Mathematics); *Sampling; *Sex Differences; *Test Items
IDENTIFIERS *Mantel Haenszel Procedure

ABSTRACT

Recent comparisons (R. Freedle and I. Kostin, 1988) of differential item difficulty (DIF) have found that items that are in some way easier differentially give advantage to white examinees. In this study, the male/female ratio was manipulated among college students taking a social sciences examination to determine the effect of the male/female ratio on the relationship between item difficulty and the male/female DIF. A total of 1,009 students comprised the male sample; the female sample included 827 students. Mantel-Haenszel delta values and associated standard error values were computed for the first 251 and 110 females in the test file as well as for the total group of females. The relationship between item difficulty and the male/female DIF appears to be very complex. Some easier items were differentially difficult for males, and others were differentially difficult for females. Larger male/female sample ratios were associated with a decrease in item difficulty among items that disadvantage women. It is cautioned that the Mantel-Haenszel statistics, although affected by the examinee variables of this study, are nonetheless sensitive to real performance differences among populations, and that care must be taken to ensure that the samples on which they are based accurately represent the examinee population. Six tables and one graph present study results. (Author/SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

GERALD E. DENAURO

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Effects of Representation of Gender
Groups in the Examinee Population
on the Mantel-Haenszel Procedure

Gerald E. DeMauro
Educational Testing Service
April 12, 1990

Paper presented at the annual meeting of the American
Educational Research Association, Boston, MA
April 1990

ED318747

TMO14701

Abstract

Recent comparisons of differential item difficulty (DIF) have found that items that are in some way easier differentially advantage White examinees (Freedle and Kostin, 1988). The current study manipulates the male-female ratio among examinees on a social sciences test to determine the effect of the male-female ratio on the relationship between item difficulty and male-female DIF. The relationship appears to be most complex. Some easier items are differentially difficult for males and others are differentially difficult for females. Larger male-female sample ratios are associated with a decrease in item difficulty among items that disadvantage women. It is cautioned that the Mantel-Haenszel statistics, though affected by the examinee variables considered in this study, are nonetheless sensitive to real performance differences among populations, and that care must be taken to insure that the samples on which they are based accurately represent the examinee population.

Effects of Representation of Gender Groups in the
Examinee Population on the Mantel-Haenszel Procedure.

Gerald E. DeMauro
Educational Testing Service

The Mantel-Haenszel procedure has been used to screen test items for differential item difficulty (DIF) at Educational Testing Service (ETS) since the Fall of 1986, when it was used to evaluate NTE Core Battery items. Earlier research exploring its application is well documented (Holland and Thayer, 1986). The procedure serves to screen both pretest items and items on final forms, after administration, but before equating and scoring.

The method addresses differential item difficulty (DIF) for various population subgroups. It provides a cumulative statistic of the log odds ratio of passing or failing an item, for two population groups first matched in overall ability. The groups are termed "reference" and "focal". Focal groups are generally composed of examinees who identify themselves as Black or African American, Hispanic, Asian American, or women. Comparisons are made with reference group members if the numbers of examinees identifying themselves by race as members of any of these groups are large enough.¹ Reference groups are composed of examinees who identify themselves as White (for Black, Hispanic, or Asian American comparisons) or by gender as men.

For examinees attaining each possible score on a test or homogenous subtest, a separate contingency table is prepared for each item as follows

		Item Performance		Total
		Pass	Fail	
Group:	Reference	W	X	N _r
	Focal	Y	Z	N _f

The odds ratios are calculated as follows

$$\chi^2_{MH} = \frac{\sum_{g=1}^G W_g Z_g / T_g}{\sum_{g=1}^G X_g Y_g / T_g} \quad (1)$$

in which g is the number of matched performance groups (usually the range of possible scores) and T_g is $N_r + N_f$. This odds ratio is then converted to the delta scale used at ETS for item difficulty:

$$\Delta_{MH} = -2.35 \ln (\chi^2_{MH}). \quad (2)$$

¹ 200 in the focal group, 600 in all for preliminary item analysis; 500 in the smaller group for test analysis; 100 in the smaller group and 500 in all for test assembly from pretested or older items.

Based on this Δ_{MH} scale and associated standard errors, items with large negative or positive DIF values, at or exceeding an absolute value of one, are reviewed by content experts for their representativeness of the desired content domain. Large positive values indicate that the item is differentially easier for the focal group (e.g., women). Large negative values indicate that the item differentially advantages the reference group (e.g., men). If items attain absolute Δ_{MH} values of 1.5 or more and are significantly greater (based on standard errors) than 1.0, they are classified as "C" items; if they attain absolutely Δ_{MH} values 1 to 1.5 and are significantly greater (based on standard errors) than either zero or than 1.0, they are classified "B"; if they attain absolute Δ_{MH} values of less than 1, they are classified as "A" items. "C" items are only used or scored under the extreme condition of having no other items from the required domain with more favorable DIF values. "B" items are only used or scored if no "A" items exist for this content domain or to assure that sufficient numbers of "A" items are available for multiple forms. "A" items may be used or scored as needed by the content requirements of the test. Naturally, all items must pass all other ETS statistical and content reviews, which are not described in this paper.

Note that the value of α_{MH} is driven by two sources of differences: between the proportions of correct answers achieved by each of the matched groups and between the proportional representation of the two groups in each of the matched groups. For example, equation 1 may be rewritten as follows:

$$\alpha_{MH} = \frac{\sum_{g=1}^G W_g (1 - P_{(F)g}) (1 - R_g)}{\sum_{g=1}^G X_g (P_{(F)g}) (1 - R_g)} \quad (3)$$

$$\text{or as: } \alpha_{MH} = \frac{\sum_{g=1}^G Z_g (P_{(A)g}) R_g}{\sum_{g=1}^G Y_g (1 - P_{(R)g}) R_g} \quad (4)$$

where: (P_f) = the proportion;
 $P_{(R)}$ = the proportion of the reference group that responds correctly;
 R = the ratio of reference group examinees to total examinees

It is clear from (3) that the influence of the performance of particular matched focal groups ($P(F)$) is largest in the scoring range where the representation of the reference group is smallest. Also from (4), the influence of the performance of particular matched reference groups ($P(R)$) increases as their representation increases. In the matched groups associated with scoring ranges where focal groups find the item easy, the Δ_{MH} will decrease sharply if there are relatively few reference group members, and the Δ_{MH} will increase. In the matched groups associated with scoring ranges where reference groups find the item easy, the Δ_{MH} will increase sharply if there are relatively many reference group members, and the Δ_{MH} will decrease.

Note too, that in each of the matched groups, as the proportion of examinees who are in the reference group approaches 1, the $P(R)$ approaches the total group P value ($P(+)$).

$$P(+) = R \cdot P(R) + (1-R) \cdot (P(F)) \quad (5)$$

The value that each of the matched groups contributes to the Δ_{MH} numerator and denominator is directly related to the value of R for that component. Finally, from (3) and (4), it is clear that as the proportion of correct responses in focal or reference groups approaches .5, the limit of α_{MH} approaches 1, and that Δ_{MH} approaches 0. In a sense, the value of the Δ_{MH} statistic is related to how disproportionately the $P(R)$ and $P(F)$ values are distributed in each of the matched groups about .5.

Research Hypotheses

In instances where the numbers of reference group examinees are relatively large in all of the matched groups, such as the case in tests that have small numbers of focal group members, then as item difficulty decreases (that is, as $P(+)$ increases), the Mantel-Haenszel delta (Δ_{MH}) value will also decrease. In such cases, Δ_{MH} may decrease as the general difficulty of the question decreases, because there may be a large ratio of reference to focal group members, and it is the reference group's performance on a such test which disproportionately determines the item difficulty.

The Mantel-Haenszel standard error values should also be related to the ratio of reference to focal group members. Standard errors should be large where the numbers of focal group members are small. Since standard error values are also used to identify high DIF items, the effects of sample reference-focal group representation and item difficulty on standard error will also be examined.

It is important to keep in mind that this study is concerned only with varying the overall reference group to focal group ratio to affect the internal ratios of each of the matched groups(g). The performances of the reference and the focal groups will not be intentionally manipulated, since the presentation of these groups at each score level would vary considerably if these two groups are varied in ability.

Research Findings

Research has commonly focused on the relationship between item content characteristics and DIF, rather than between item difficulty and DIF. Content characteristics that reduce item difficulty should also decrease (make more negative) Mantel-Haenszel delta values for tests having relatively small focal populations. Generally, this is more likely to be observed in ethnic comparisons where the majority of examinees are often White, than in male-female comparisons, which generally involve similar numbers of reference and focal group members.

In fact, Freedle and Kostin (1988) found that White examinees perform differentially better than Black or African American examinees on GRE verbal analogy items that contain the most commonly used vocabulary. Using log linear chi-square procedures which rely on similar matching techniques as the Mantel-Haenszel, Scheuneman (1985) found that:

- 1.) White examinees are advantaged relative to matched groups of Black examinees when items employ commonly used vocabulary; and that
- 2.) Cues imbedded in items also differentially advantage White examinees relative to matched groups of Black examinees.

In each of the findings, the content characteristics associated with negative Δ_{MH} values are those that make the item easier. The findings appear to support the hypothesis that differentially difficult items should be easier, because easy items in tests taken by disproportionately large reference groups, are those items that the reference group finds easy.

Male-Female Comparisons

Secondary analyses of male-female comparisons also typically explore content characteristics that differentially advantage one group or the other (McPeck and Wild, 1987). However, as yet, research on item difficulty and Mantel-Haenszel analysis has been limited.

Among achievement tests, the ratios of male and female examinees vary as a reflection of differential interest in various fields. An achievement test in the social sciences was examined in this study because it is typically taken by roughly equal numbers of male and female examinees, who can be expected to show reasonably similar attraction to the field. The study is concerned with how manipulations of the size of the focal population (female) would affect the Δ_{MH} and Mantel-Haenszel standard error values of items. It is hypothesized that the effects of this manipulation would interact with item difficulty, and, in particular that item difficulty and Δ_{MH} values would be related to a greater extent in samples with relatively smaller focal groups.

Methods

Test. The test from which the items were drawn normally has 150 scored items. However, only the first 44 items¹ were used in this study, to enable generalization to shorter instruments or instances in which subtest scores are used as the Mantel-Haenszel matching criterion. This reduction in instrument length, while desirable for the reasons given, resulted in a loss of reliability. Coefficient alpha was computed to be .797.

Samples. The male sample (n=1009) was left intact. The female sample was originally 827 (45.0 percent), Mantel-Haenszel delta values and associated standard error values were computed for both the first 251 (19.9 percent) and 110 (9.8 percent) females in the test file as well as for the total group. The values were then compared.

The means and standard deviations of the three samples, respectively labelled "55-45", "80-20", and "90-10" are presented in Table 1. The lowest female mean was attained in the "80-20" sample. Comparing this group to the 576 females who were not members of the "80-20" sample (mean=23.48) shows that the samples were fairly homogenous in skills ($t(df=825) = 1.335, ns$).

Stability of construct. Table 2 shows delta correlations of the 44 items among the three samples. This correlation is generally taken as a measure of the homogeneity of the measure over population groups (Angoff and Modu, 1973). The delta scale measures item difficulty and is a normal deviate transformation of the P(+) scale, having a mean of 13 and a standard deviation of 4, with higher P(+) values yielding lower delta values.

The homogeneity of the construct across the different samples is stable (.999, .997, .906). By extension of the logic behind the delta correlations, the correlation of Mantel-Haenszel delta values should suggest how homogenous the differences among males and females are from sample to sample. These correlations should be somewhat lower in magnitude than the item delta correlations if the sample male-female ratio interacts with item difficulty in its effect on Mantel-Haenszel delta (Δ_{MH}) values. In fact, the Mantel-Haenszel delta correlations are .906, .851 and .758. Thus, the relative differential difficulty of particular items has not been dramatically affected by the sampling procedures.

¹ Actually, the first 45 items were drawn, but one was inadvertently scored incorrectly for the study and was subsequently eliminated.

Mantel-Haenszel Delta Differences

One characteristic of the odds ratio (Δ_{MH}) is that it will sum to approximately one over the total number of test items that comprise the group matching criterion. This can be shown by considering that, over all items, matched groups must get the same number of items correct. If omitted and not reached items were counted as wrong, and the analyses were based on the same number of both focal and reference group members throughout, the sum of Mantel-Haenszel delta values would be zero. In practice examinees who do not reach an item are excluded from the analysis of the item.

DeMauro and Olson (1989) have shown that Δ_{MH} values have larger variance and greater sensitivity to small differences in performance between focal and reference groups when not reaching examinees are included in the analyses. Since, in the artificial test constructed for this study, not reaching examinees are not found, and since the current analysis intends to be most conservative to discern only real effects that might exist, omitted responses were excluded from the analysis.

As this is a rights-scored test, with no penalty for guessing, examinees who omitted items were not included in analyses to avoid the introduction of dimensions related to omission preferences in the matching criterion. Earlier research has found that such introductions of multidimensionality in the matching criteria results in a less conservative test (Bleistein and Schmitt, 1987; Dorans, Schmitt and Bleistein, 1988).

In practice, the Mantel-Haenszel analyses are conducted twice. Following the suggestions of Holland and Thayer (1986), items with particularly high absolute Δ_{MH} values, type "C" items, are removed from the matching criterion. Analysis is then made using groups matched on the purified criterion. In the current study there were no "C" items for any of the three samples, and the results in fact represent a final analysis.

Because the Δ_{MH} means must approximate zero under all sample conditions, this study is more concerned with the effects of the interaction of item difficulty and sample ratio on Δ_{MH} than with the main effects of sample ratio. It is quite possible that in samples having relatively large reference groups, increases in the number of large negative Δ_{MH} items will be associated with an increase in the number of small positive Δ_{MH} items or large positive Δ_{MH} items, in accordance with the characteristic that the Δ_{MH} values will sum to zero within samples. More importantly, the study focuses on the changing relationship between item difficulty and Δ_{MH} as sample ratios change.

Analysis

Items were first divided by general level of difficulty (delta values) into four groups (Table 2). These groupings, having 13, 10, 10 and 11 items, respectively, were designed so that the range of the item difficulty delta values within groups did not overlap with delta values from any of the same or other samples. For example, the easiest of the group of hard items in any of the samples had a higher delta value than the hardest item in any of the other group of items of the three sample ratios. Table 3 shows the means and standard deviations of these groupings.

In the first set of analyses, a mixed analysis of variance was used to evaluate the effect of sample ratio and item difficulty on Δ_{MH} and Mantel-Haenszel standard error values. Group membership was the between items effect (different groups of items having different values). The within items effect (the sample item having different values) was sample ratio: 55-45, 80-20, 90-100.

To test whether low Δ_{MH} values were more likely to be associated with easier items in the 90-10 sample or in the 80-20 sample than in the 55-45 sample, a second analysis of variance was computed, in which items were assigned to three categories based Δ_{MH} values in each sample. The first category contained all items having negative Δ_{MH} values in each sample. The second category contained all items having positive Δ_{MH} values in each sample, and the third category contained all other items. A mixed between items and within items analysis of variance was then computed, in which the within items effect was male-female sample ratio, and the between items effect was the Mantel-Haenszel item category described above. Tukey post hoc contrasts for the sample ratio effect (equal ns) and Scheffe contrasts within sample ratios (unequal ns) (Myers, 1973). Table 3 also shows the mean Δ_{MH} values for the items in each of the categories.

Finally, regression analyses were conducted of standard errors and Δ_{MH} values onto item difficulty (deltas). These general linear models were used to discern the magnitude of relationships among these variables.

Results

Mantel-Haenszel Delta Values (Tables 4 and 5).

As expected, the Mantel-Haenszel delta values did not change in relation to sample ratios ($F(df=2, 80)=1.94, ns$). Unexpectedly, they did not change, either, as a function of the interaction of sample ratio and item difficulty ($F(df=6, 80)=1.42, ns$) (Table 4).

In the second analysis, in which item difficulty was the dependent variable, significant effects were observed both for sample ratio ($F(df=2, 82)=44.73, p<.001$) and the interaction of Δ_{MH} and sample ratio ($F(df=4, 82)=9.64, p<.001$). Tukey post hoc contrasts revealed that the mean item difficulty values on the delta scale for the 55-45 sample (11.69) were significantly higher than for the 80-20 sample (11.56), which were, in turn, significantly higher than for the 90-10 sample (11.51). This is a natural consequence of reducing the size of the lower scoring group. Post hoc contrasts also showed that among consistently negative Δ_{MH} items, the item difficulty values went down significantly as the male-female sample ratio declined (11.31, 11.08, 10.96, respectively for the 55-45, 80-20, and 90-10 samples). For the consistently positive Δ_{MH} items, there were no differences in item difficulty values in relation to sample ratios (12.48, 12.43, 12.44, respectively). Finally, for the sometimes negative Δ_{MH} values, the mean difficulty for the full sample (11.17) was significantly higher than for the 80-20 (11.04) or the 90-10 (10.99) samples (Figure 1). The effects of reducing the proportion of the lower scoring group is most pronounced on negative DIF items.

Mantel-Haenszel Standard Errors (Table 4)

Significant effects were observed for item difficulty ($F(df=3, 40)=5.81, p<.01$), sample ratio ($F(df=2, 80) = 19.87, p<.001$), and the interaction of these variables ($F(df=6, 80)=5.45, p<.001$). Tukey post hoc contrasts showed that the mean standard error of the 90-10 sample (1.14) was significantly higher than that of the 80-20 sample (0.79), which, in turn, was significantly higher than that of the 55-45 sample (0.54). Also, while the mean standard error is largest among the easiest items, it is particularly large for easy items in the 90-10 sample.

Regression of Δ_{MH} and Standard Error on Item Difficulty (Table 6).

Table 6 shows that the regression coefficients of standard error and Δ_{MH} on item difficulty varied in relation to whether the Δ_{MH} values were negative or positive. The curvilinear relationship between Δ_{MH} and item difficulty, in particular, may explain why there was a failure to discern the impact of item difficulty on Δ_{MH} values, earlier.

Discussion

Interpretation

the Δ_{MH} . Two seemingly disparate findings are reported: (1) the Δ_{MH} values are not reliably influenced by the interaction of item difficulty and size of the focal group; and (2) For items having negative Mantel-Haenszel values in all samples, sample size affects item difficulty values.

Perhaps some of this is explained by the regression of Δ_{MH} values and Mantel-Haenszel standard error values onto item difficulty. The Δ_{MH} coefficients demonstrate a curvilinear relationship between item difficulty and Δ_{MH} values which may have masked the interaction of item difficulty and sample ratio in the first analysis. For positive Δ_{MH} values, the relationship is negative, and for negative values, the relationship is positive. As item difficulty decreases, some items become differentially more difficult for the focal group and others become differentially more difficult for the reference group. Also, among the three sample ratios, the effect is only reliable for positive items in the 55-45 sample. From equation 3, it is clear that if Y_i , the number of focal group members that answer the item correctly, increases, either because the size of the focal group increases, or the probability of a correct response from the focal group increases (equation 2), or both, the value of α_{MH} will decrease and the value of Δ_{MH} will rise. This explains why, in the sample with the highest female representation, Δ_{MH} increased as item difficulty decreased. The effect of changes in Y_i were obviously less pronounced for groups with fewer females.

This also explains the Δ_{MH} by sample ratio interactive effect on item difficulty in the second analysis. From equation 4, it is clear that where women have greater representation, the Y_i values contribute more heavily to the relationship of Δ_{MH} values and item difficulty while where they have less representation, the $P_{(R)}$ values contribute more both to Δ_{MH} values and to item difficulty. As the number of women who answer an item correctly (Y_i) exceeds the number who answer an item incorrectly (Z_i), Δ_{MH} becomes increasingly positive. If the representation of females is high, the item difficulty drops and the relationship is negative, as observed most clearly for positive Δ_{MH} items in the high 55-45 group.

Among negative Δ_{MH} items, $P_{(R)}$ may be greater than .5, and as $P_{(R)}$ increases, the item difficulty decreases, especially in samples in which the small numbers of women, do not have as great an impact on item difficulty. Hence, item difficulties of the 90-10 sample's item were significantly lower than those of the other samples for the negative Δ_{MH} items, and positive relationships between item difficulty and Δ_{MH} were observed. Item difficulty may drop, then, because males or females, or both answer the item correctly. It's relationship to Δ_{MH} values, however, depends on the relative representation of the group, as well as performance.

Grouping items by difficulty, as was done in the first analysis, was a poor technique to observe these relationships. Among easy items, Δ_{MH} is likely to be low in the 90-10 group, but it could also be low in the 55-45 group. It was expected that easier items would be disproportionately found among negative Δ_{MH} items. However, that does not imply that negative Δ_{MH} items would be disproportionately found among easier items, since negative Δ_{MH} items could arise from relatively poor female performance or relatively good male performance.

Standard errors. The reliably negative relationships between item difficulties and the standard errors are also interesting. Either for increasingly larger samples (55-45 and 80-20), or for increasingly lower male-female ratios, among negative Δ_{MH} items, more difficult items are associated with lower standard errors. Among negative Δ_{MH} items, lower Δ_{MH} values are associated with larger standard error values. Taken together, these findings suggest that lower negative Δ_{MH} items are more likely to be taken out of the test or not scored when samples are larger, since the lower standard errors will make these Δ_{MH} values significant.

Conversely, in smaller samples, lower negative Δ_{MH} values may not be rejected. The ANOVA results suggest that smaller sample size or larger sample ratio may inflate standard errors especially for easier items. It follows, therefore, that while the negative Δ_{MH} values may decrease, in large male-female sample ratios, the likelihood that these items will be identified as problem items is offset by the increase in the standard error. It remains to determine if this increase is a function of sample ratio increase or sample size decrease.

Implications

The manipulations of male-female sample ratios and subsequent effects on the Mantel-Haenszel statistics raises the question of how to attain unbiased estimates of population parameters. The answer is that the samples drawn for computation of the statistics must approximate the population in (1) ability of both the focal (female) and reference (male) groups, (2) overall population representation, and (3) representation among ability groups. This could be accomplished by randomly drawing samples (as is currently done at ETS) without stratification on group membership or ability.

It should be cautioned, that test development attempts to equalize population group performance on items by reducing item difficulty may backfire, particularly when the focal group is relatively small. Such attempts may well raise the probability of differential item difficulty.

Finally, it must be particularly emphasized that none of the manipulations of this study affect the finding that high absolute Mantel-Haenszel delta values are a valid index of differential item difficulty. Rather, the study explores effects of sampling and possible explanations for earlier findings among ethnic comparisons this time using groups that are usually equally represented in examinee groups.

References

- Angoff, W.H., & Modu, C.C. (1973). Equating the Scales of the Prueba de Aptitud Académica and the Scholastic Aptitude Test - Research Report 3. New York: College Entrance Examination Board.
- Bleistein, (A. & Schmitt, A.P. (1987). The Effects of Criterion Selection on the Evaluation of Differential Item Functioning on SAT Analogy Items between Black and White Examinees. Paper presented at the annual meeting of the National Council on Measurement in Education Washington, D.C.
- DeMauro, G.E. & Olson, J.F. (1989). The Impact of Differential Speededness on DIF. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Dorans, N.J, Schmitt, A.P. & Bleistein, C.A. (1988). The Standardization Approach to Assessing Differential Speededness. Research Report-88-31. Princeton, NJ: Educational Testing Service.
- Freedle, R. & Kostin, I. (1988). Relationship between item characteristics and an index of differential item functioning for the four GRE verbal item types. Princeton, NJ: GRE Board Professional Report.
- Holland, P.W. & Thayer, D.T. (1986). Differential Item Functioning and the Mantel Haenszel Procedure. Program Statistics Research Technical Report No. 86-69, Research Report No. 86-31. Princeton, NJ: Educational Testing Service.
- McPeck, W.M. & Wild, C.L. (1987). Characteristics of Quantitative Items that Function Differently for Men and Women. Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.
- Myers, J.L. (1973). Fundamentals of Experiment Design (2nd Edition). Boston, MA.: Allen, Bacon, Inc.
- Scheuneman, J.D. (1985). Exploration of causes of bias in test items. Princeton, NJ: Graduate Record Examinations Board Professional Report GREB No. 81-21 P. ETS Research Report 85-42.

Figure 1
Mean Item Difficulty
In Relation to Mantel-
Haenszel Delta (Δ_{MH})

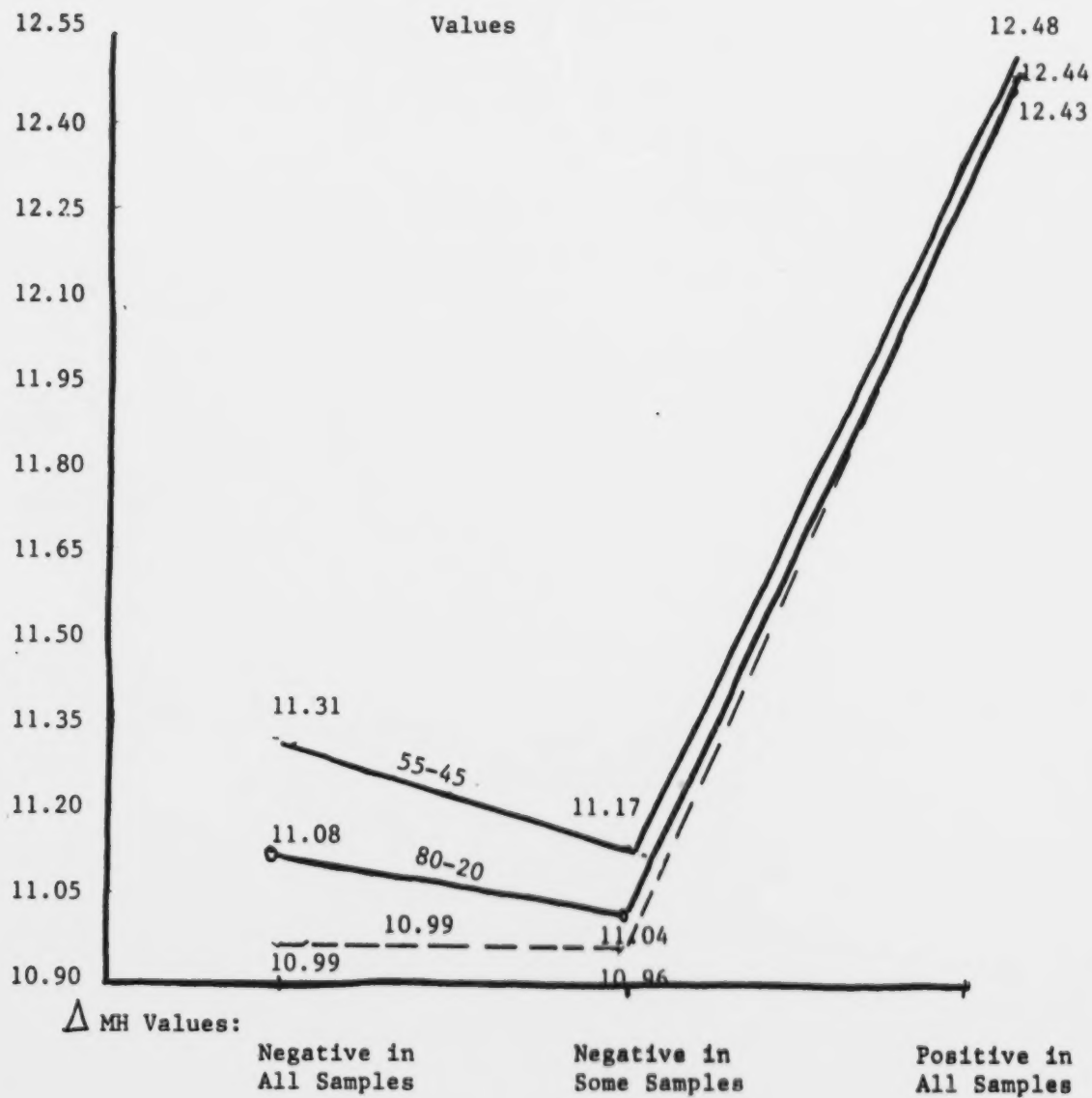


Table 1

Means and Standard Deviations for
Males and Females on 44 Item Test,
by Sample Ratios
(Males - Females)

<u>Sex</u>	<u>n</u>	<u>Ratio</u>	<u>Mean</u>	<u>Standard Deviations</u>
Females	827	55-45	23.303	5.787
	251	80-20	22.896	5.796
	111	90-10	23.127	5.685
Males	1009	All	25.289	5.625

Table 2

Correlations of Item Difficulty
(Delta) and Mantel-Haenszel DIF
values (on the Scale of Delta
Differences) for 44 Items Drawn
from Samples Having Different
Numbers of Females

		Sample Ratio (Male-Female)					
<u>Sample Ratios</u>		55-45		80-20		90-10	
		Δ	Δ_{MH}	Δ	Δ_{MH}	Δ	Δ_{MH}
55-45	Δ	-	.021	.999	-.042	.997	-.133
	Δ_{MH}	-		.065	.906	.082	.758
80-20	Δ	-	-	-	-.011	.999	-.106
	Δ_{MH}	-	-	-	-	.005	.851
90-10	Δ	-	-	-	-	-	-.097
	Δ_{MH}	-	-	-	-	-	-
Means		11.686	-0.043	11.561	.042	11.506	.092
S.D.		2.786	0.555	2.811	0.708	2.831	.840

Table 3

Item Difficulty and Item Mantel Haenszel
Delta Means and Standard Deviations,
by Groupings used in Study Analyses

First Grouping: by Difficulty

		Sample: 55-45		80-20		90-10	
		Delta		Delta		Delta	
<u>Group</u>	<u>N</u>	<u>Mean</u>	<u>S.D.</u>	<u>Mean</u>	<u>S.D.</u>	<u>Mean</u>	<u>S.D.</u>
Easiest	11	7.97	1.68	7.78	1.62	7.68	1.58
Next Easiest	10	11.01	0.35	10.89	0.29	10.83	0.27
Next Hardest	10	12.39	0.61	12.30	0.56	12.27	0.55
Hardest	13	14.81	1.00	14.72	1.04	14.67	1.11

Second Grouping: by Δ_{MH}

		Sample: 55-45		80-20		90-10	
		Δ_{MH}		Δ_{MH}		Δ_{MH}	
<u>Group</u>	<u>N</u>	<u>Mean</u>	<u>S.D.</u>	<u>Mean</u>	<u>S.D.</u>	<u>Mean</u>	<u>S.D.</u>
All negative	14	-0.57	-0.41	-0.69	0.47	-0.72	0.35
Some negative	14	-0.06	0.27	0.03	0.26	0.09	0.38
All positive	16	0.43	0.41	0.69	0.49	0.80	0.81

Table 4

Mean Mantel-Haenszel Delta
and Standard Error Values
in Relation to Male-Female Sample Ratio

<u>Sample Ratio</u>	<u>Item Difficulty</u>	<u>Δ_{MH}</u>	<u>Mantel Haenszel Standard Error</u>
55-45	Easiest	-0.17	0.74
	Next Easiest	-0.04	0.46
	Next Hardest	-0.15	0.43
	Hardest	-0.08	0.50
	All	-0.04	0.54
80-20	Easiest	0.07	1.09
	Next Easiest	-0.07	0.68
	Next Hardest	0.23	0.65
	Hardest	-0.05	0.73
	All	0.04	0.79
90-10	Easiest	0.27	1.63
	Next Easiest	-0.13	0.94
	Next Hardest	0.22	0.90
	Hardest	0.01	1.07
	All	0.09	1.14

Table 5

Item Difficulty Means in Relation to
Male-Female Sample Ratio and Mantel-Haenszel
Delta Values (Means sharing the same Underline
are Not Significantly Different)

Mantel-Haenszel Delta Values	Male-Female Sample Ratio		
	<u>55-45</u>	<u>80-20</u>	<u>90-10</u>
Negative in All samples (n=14)	11.31	11.08	10.96
Negative in One or two Samples (n=16)	11.17	<u>11.04</u>	<u>10.99</u>
Positive in All Samples (n=14)	<u>12.48</u>	<u>12.43</u>	<u>12.44</u>

Table 6

General Linear Regression:
 Δ_{MH} and Mantel-Haenszel
 Standard Error Values onto
 Item Difficulty, by Male-Female
 Sample Ratio and Whether Δ_{MH}
 Values are Positive or Negative

Male-Female Sample Ratio	Δ_{MH} Positive or Negative	No. Items	Standard Error Coefficient	Δ_{MH} Coefficient	Intercept	R-Square
55-45	Neg.	22	-6.58***	0.43	15.02	.433
	Pos.	22	1.17	-3.33*	12.99	.271
	All	44	-6.82***	0.09	15.34	.319
80-20	Neg.	20	-4.56***	0.19	14.86	.489
	Pos.	24	-7.31	-0.80	18.06	.259
	All	44	-5.13***	0.18	15.60	.341
90-10	Neg.	19	-2.92	1.11	15.02	.073
	Pos.	25	-2.99***	-0.33	15.49	.440
	All	44	-3.08**	0.37	14.98	.299

* Exceeds the $p < .05$ level of significance (coefficient > 0)

** Exceeds the $p < .01$ level of significance (coefficient > 0)

*** Exceeds the $p < .001$ level of significance (coefficient > 0)